MAG-Edit: Localized Image Editing in Complex Scenarios via Mask-Based Attention-Adjusted Guidance

Qi Mao[1] Lan Chen[1] Yuchao Gu[2] Zhen Fang[1] Mike Zheng Shou[2]

[1]Communication University of China [2]Show Lab, National University of Singapore

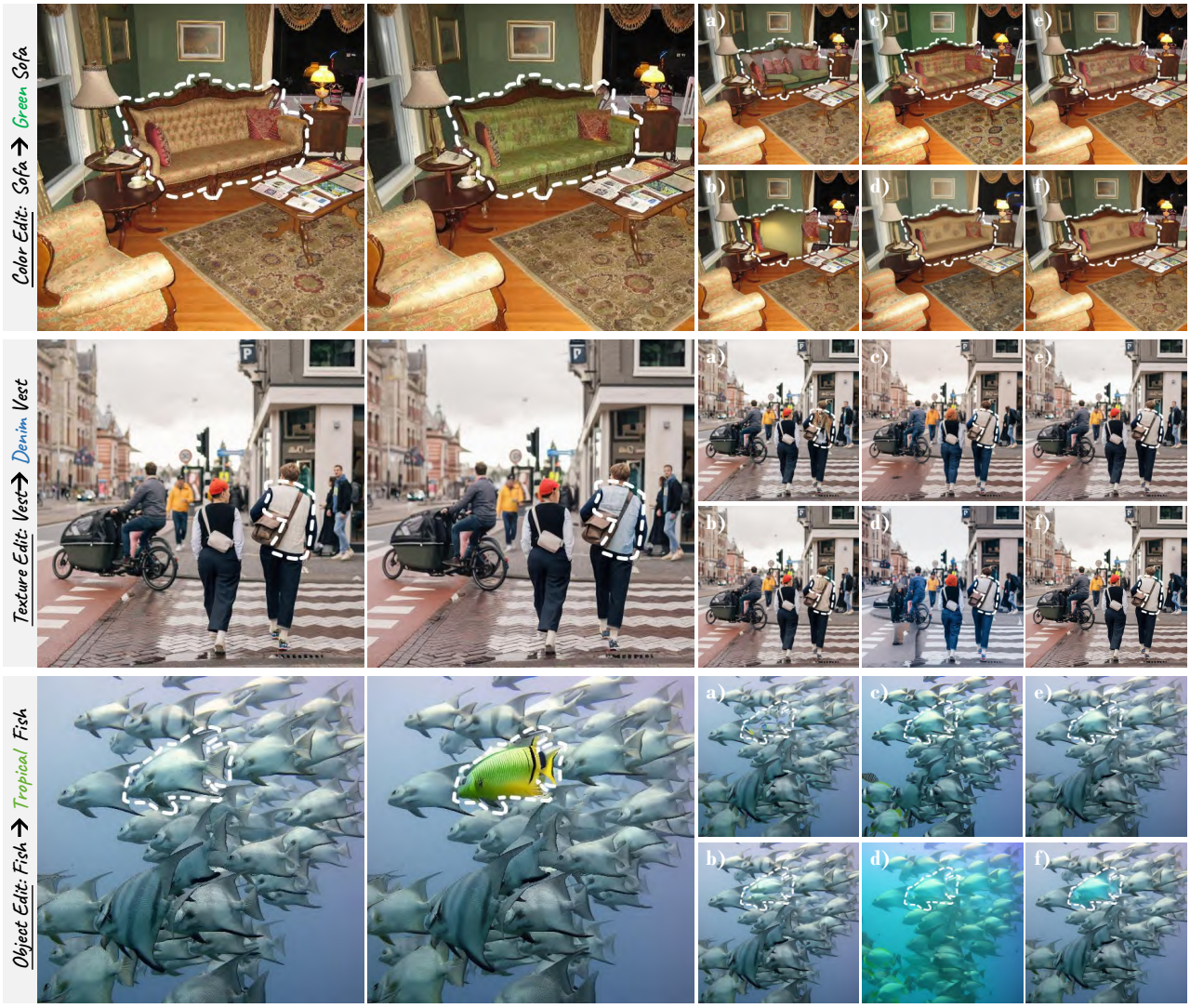


Figure 1. **Localized image editing in complex scenarios.** We enhance the visibility of the editing regions by outlining their edges with white dashed lines. Existing mask-based approaches, *e.g.*, (a) Blended latent diffusion [2] and (b) DiffEdit [8], often modify the structural details of the original edited areas, causing significant discrepancies with their surrounding context. Mask-free attention-based methods, such as (c) Prompt2Prompt (P2P) [10] and (d) Plug-and-play (PnP) [28], exhibit leakage, becoming aligned with incorrect regions, resulting in inefficiencies in prospective areas. By incorporating blending operations, (e) P2P+Blend and (f) PnP+Blend can mitigate leakage; however, their editing effect in the prospective region still lacks efficiency. In contrast, our method **MAG-Edit** achieves localized image edits that align well with the target prompt and maintain structural integrity more effectively.

Abstract

*Recent diffusion-based image editing approaches have exhibited impressive editing capabilities in images with simple compositions. However, localized editing in complex scenarios has not been well-studied in the literature, despite its growing real-world demands. Existing mask-based inpainting methods fall short of retaining the underlying structure within the edit region. Meanwhile, mask-free attention-based methods often exhibit editing leakage and misalignment in more complex compositions. In this work, we develop **MAG-Edit**, a training-free, inference-stage optimization method, which enables localized image editing in complex scenarios. In particular, MAG-Edit optimizes the noise latent feature in diffusion models by maximizing two mask-based cross-attention constraints of the edit token, which in turn gradually enhances the local alignment with the desired prompt. Extensive quantitative and qualitative experiments demonstrate the effectiveness of our method in achieving both text alignment and structure preservation for localized editing within complex scenarios.*

1. Introduction

Text-based image editing aims to manipulate images in accordance with provided textual prompts. Recent advancements in large-scale text-to-image (T2I) diffusion models, such as Stable Diffusion [23], DALL·E [22], and Imagen [25], have demonstrated remarkable ability to generate high-quality, diverse images that accurately reflect specified textual descriptions. Trained on comprehensive datasets, these models effectively connect textual descriptions with corresponding images, thereby paving a new way for text-based image editing.

The past year has witnessed a substantial increase in the development of methods using diffusion models for text-based image editing, which can be broadly categorized into three groups: training [3, 31], fine-tuning [14, 24, 32], and training-free methods [4, 8, 10, 18, 20, 28]. Existing approaches predominantly concentrate on manipulating *prominent* objects within *simple* compositions. However, images in real-world scenarios usually contain intricate compositions with multiple objects. Additionally, users often require edits in specific localized regions. For example, in home interior design, a user might wish to change the color of a particular piece of furniture to better complement the surrounding space. A case in point is altering the color of a sofa to green, as illustrated in the first row of Fig. 1, to improve its aesthetic coherence with the environment.

The trade-off between *fidelity* and *editability* in localized image editing within complex scenarios presents significant challenges. Mask-based inpainting methods directly generate a new object as a foreground element and blend it into the original image [1, 2, 8, 12, 29]. However, this often results in substantial structural changes within the edited areas, causing noticeable discordance with their complex surroundings, as shown in the third column of Fig. 1. On the other hand, mask-free methods that utilize attention injection mechanisms such as Prompt-to-Prompt (P2P) [10] and Plug-and-Play (PnP) [28] can preserve the original image's structure and layout. Nevertheless, they struggle to precisely align the local editing region with the intended text in intricate scenarios, largely due to their reliance on the text prompts' localization capabilities. As a result, editing effects often extend beyond the intended area and impact incorrect regions, as shown in the fourth column of Fig. 1. Integrating mask-based blending techniques into P2P and PnP can alleviate leakage, but issues with incorrect alignment remain unresolved. This misalignment leads to the absence of editing effects in the intended areas, demonstrated in the last column of Fig. 1.

In this work, we introduce a novel editing scheme named ***M**ask-Based **A**ttention-Adjusted **G**uidance* (**MAG**-Edit). This approach is designed to enable *localized* image editing in *complex* scenarios, which typically involve intricate compositions with multiple objects. Given that cross-attention (CA) maps in pre-trained T2I diffusion models effectively capture the correlation between input features and text embeddings, our key insight is that *adjusting the noise latent feature to attain higher CA values significantly enhances its alignment with the corresponding text prompt*. As a result, we propose locally optimizing the noise latent feature during the inference stage by maximizing two distinct mask-based CA constraints tailored for the target editing prompt. In particular, our approach aims to maximize two aspects of ratios: first, the CA value of the edit token in relation to all token CA values within the masked area, and second, the CA value of the edit token inside the mask compared to its overall CA values. Subsequently, the gradients of these constraints guide the update of the noise latent feature, thus progressively aligning the editing effect with the desired text prompt and spatial requirements. The effectiveness of the proposed method is evident in the second column of Fig. 1.

The main contributions of our work can be summarized as follows,

- We introduce MAG-Edit, a novel training-free, inference-stage optimization scheme. To our knowledge, this is the first method specifically designed to address localized image editing in complex scenarios.
- We propose two mask-based CA constraints in terms of the token and spatial ratio, guiding the local noise latent feature to better align with the target text.
- We extensively validate MAG-Edit's efficiency in localized image editing across diverse intricate indoor and outdoor scenarios. Quantitative and qualitative experimental results demonstrate a significantly im-

proved trade-off between editing efficiency and structure preservation when compared to existing baselines.

2. Related Work

Text-Based Image Editing Using Diffusion Models can be mainly classified into three categories: training [3, 31], fine-tuning [14, 24, 32], and training-free methods [4, 8, 10, 18, 20, 28]. InstructPix2Pix [3] requires significant resources for extensive training, while fine-tuning methods like Imagic [14] risk overfitting by optimizing the full model with limited data. In this work, we focus on training-free methods. Some approaches [1, 2, 12, 29] utilize masks to generate foreground objects and blend them into the original image through blending operations. In particular, the Blended Diffusion [1] and Blended LD [2], directly generate foreground objects based on text prompts. DiffEdit [8, 29] introduces an unsupervised method for learning the mask and employs DDIM inversion [26] noise latent features alongside the target prompt to generate the foreground image. Although these approaches successfully maintain the integrity of unedited regions outside of the mask, they may introduce large structural changes within the edit regions, causing inconsistencies with the surrounding context in complex scenes. Other methods [4, 10, 28] such as P2P [10] involve the attention integration mechanisms to maintain the structure and layout of the original image. Recent advancements in inversion methods [15, 17, 18] propose to improve DDIM inversion [26] for encoding real images, achieving improved reconstruction and more flexible editing capabilities. However, the integration of P2P [10] remains essential for these methods to facilitate image editing. When applied to localized editing in intricate scenarios, attention-based methods often result in leakage to incorrect areas, leading to inefficiencies in prospective regions.

Optimization on the Noise Latent Feature. Recent advances [5, 6, 30] in image generation with diffusion models have investigated the use of CA constraints to optimize the noise latent feature during inference. The pioneering work, Attend-and-Excite [5], addresses issues like catastrophic neglect and incorrect attribute binding by maximizing the largest CA units corresponding to all subject tokens in the text prompt. This approach refines the noise latent feature at each diffusion step, thereby guiding the model to generate all subjects described in the text accurately. Several training-free layout-generation methods [6, 30] propose to optimize the noise latent feature by maximizing CA constraints in conjunction with bounding boxes, allowing objects to appear in specific regions. While the image generation process has demonstrated effectiveness, the application of noise latent feature optimization to image editing has received relatively less attention. Pix2pix-zero [20] offers a solution by optimizing the noise latent feature, constraining the CA maps of the editing branch to align with the reconstruction branch, thus preserving the original image's structural layout. In contrast to structural preservation, the proposed method aims to align the local noise latent feature more semantically with the target text prompt, enabling localized editing in complex scenarios.

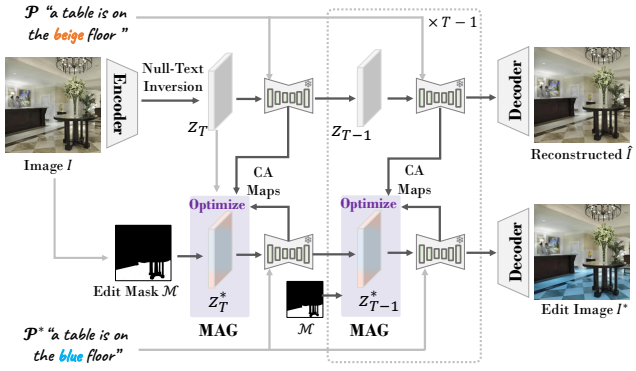


Figure 2. **High-level overview of the proposed MAG-Edit framework.** The first and second rows represent the reconstruction and editing branches, respectively. In the editing branch, the noise latent feature is optimized through MAG. This optimization process aids in achieving alignment with the target edit prompt "**blue**" within the intended edit region $\mathcal{M}$.

3. Background and Preliminaries

Stable Diffusion (SD) [23] aims to denoise the random sampled noise latent z_T conditioned on text embedding $\mathcal{C}$. This process transforms z_T into a series of noise latent features z_t at each diffusion step t, where $t \sim [1, T]$ and T is the timestep number. To train the diffusion model ε_θ, the initial latent feature z_0 undergoes an interactive process by adding Gaussian noise ε to the noise latent features z_t. Then, the network is minimized by,

$$\min_{\theta} E_{z_0, \varepsilon \sim N(0,I), t \sim \text{Uniform}(1,T)} \left\| \varepsilon - \varepsilon_\theta \left(z_t, t, \mathcal{C} \right) \right\|_2^2 . \quad (1)$$

Furthermore, the classifier-free guidance [11] performs unconditional prediction to mitigate the amplifying effect of text-based conditioning as:

$$\tilde{\varepsilon}_\theta \left(z_t, t, \mathcal{C}, \varnothing \right) = w \cdot \varepsilon_\theta \left(z_t, t, \mathcal{C} \right) + (1 - w) \cdot \varepsilon_\theta \left(z_t, t, \varnothing \right), \quad (2)$$

where $\varnothing$ is the unconditional embedding of a null text, and w is the guidance weight. To generate images from given z_T, we can employ deterministic DDIM sampling [26] as:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \tilde{\varepsilon}_\theta \left(z_t, t, \mathcal{C}, \varnothing \right). \quad (3)$$

Null-Text Inversion. Real image editing requires reversing corresponding z_0 back to z_T. A straightforward DDIM inversion method [26], in theory reversible with infinitesimally small steps, tends to accumulate reconstruction errors

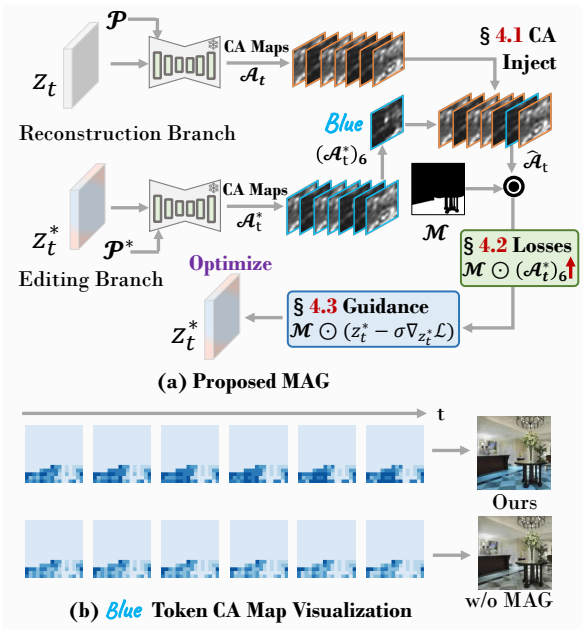


Figure 3. **Illustration of Our MAG.** (a) We optimize the z_t^* by maximizing the mask-based CA constraints of target token (*e.g.*, "**blue**"). (b) The top row and the bottom row illustrate the average mask-based CA maps of "**blue**" token with our MAG and "w/o MAG" from different diffusion steps, respectively. Applying optimization of z_t^* with MAG, there is a noticeable enhancement in the CA values within the edit regions $\mathcal{M}$.

in the denoising process, particularly due to classifier guidance. To address this, Null-text inversion [18] aligns the diffusion latent trajectory with the denoising latent trajectory by optimizing a step-wise unconditional embedding $\varnothing_t$.

Prompt-to-Prompt (P2P) [10] introduces several prompt-based editing operations leveraging CA maps: First, the *word swap* involves injecting all attention maps in the reconstruction branch, generated by the source prompt, into attention maps from the editing process using the target prompt. In contrast, the *prompt refinement* selectively replaces the CA maps associated with tokens common to both the source and target prompts. Furthermore, P2P introduces the *attention re-weighting* operation, enabling direct scale adjustments to the CA values. This technique is specifically designed to control the granularity of the editing process.

4. Methodologies

Let $\mathcal{I}$ be a real image, we first employ Null-text inversion [18] to encode it into the noise latent feature z_T. Given the original text prompt $\mathcal{P}$ and edited prompt $\mathcal{P}^*$, we define the set of new target tokens as $\mathcal{S}^* = \{s_1^*, s_i^*, ..., s_I^*\}$ present in $\mathcal{P}^*$ against $\mathcal{P}$, the common tokens as $\mathcal{S} = \{s_1, s_j, ..., s_J\}$ and $\mathcal{S}^* \cap \mathcal{S} = \emptyset$. An edit region mask $\mathcal{M}$ derived by $\mathcal{I}$ is provided to precisely localize the edit region. Fig. 2 illustrates the high-level overview of the proposed editing framework, which consists of two branches, *i.e.*, reconstruction and editing branches generated by prompt $\mathcal{P}$ and $\mathcal{P}^*$, respectively. In this work, we aim to optimize the noise latent feature z_t^* of the editing branch at diffusion step t. Our objective is to align the desired editing effects specified by $\mathcal{S}^*$ with the prospective region defined by $\mathcal{M}$, which enables localized editing in complex scenarios.

To achieve this, we first inject CA maps of common tokens similar to the *prompt refinement* in P2P [10] (Section 4.1). Subsequently, we introduce MAG to automatically manipulate z_t^*, which contains two key steps: defining two mask-based constraints in Section 4.2 and performing gradient guidance in Section 4.3.

4.1. Attention Injection

As illustrated in Fig. 3 (a), to preserve the structural information of the original image, CA maps of common tokens from the reconstruction branch are first injected into the editing branch at diffusion step t, thereby obtaining the mixing CA maps $\hat{A}_t$ as:

$$Inject(A_t, A_t^*) := \begin{cases} (A_t)_j & j \in \{1, j, \cdots, J\}, \\ (A_t^*)_i & i \in \{1, i, \cdots, I\}. \end{cases} \quad (4)$$

4.2. Mask-Based Attention-Adjusted Constraints

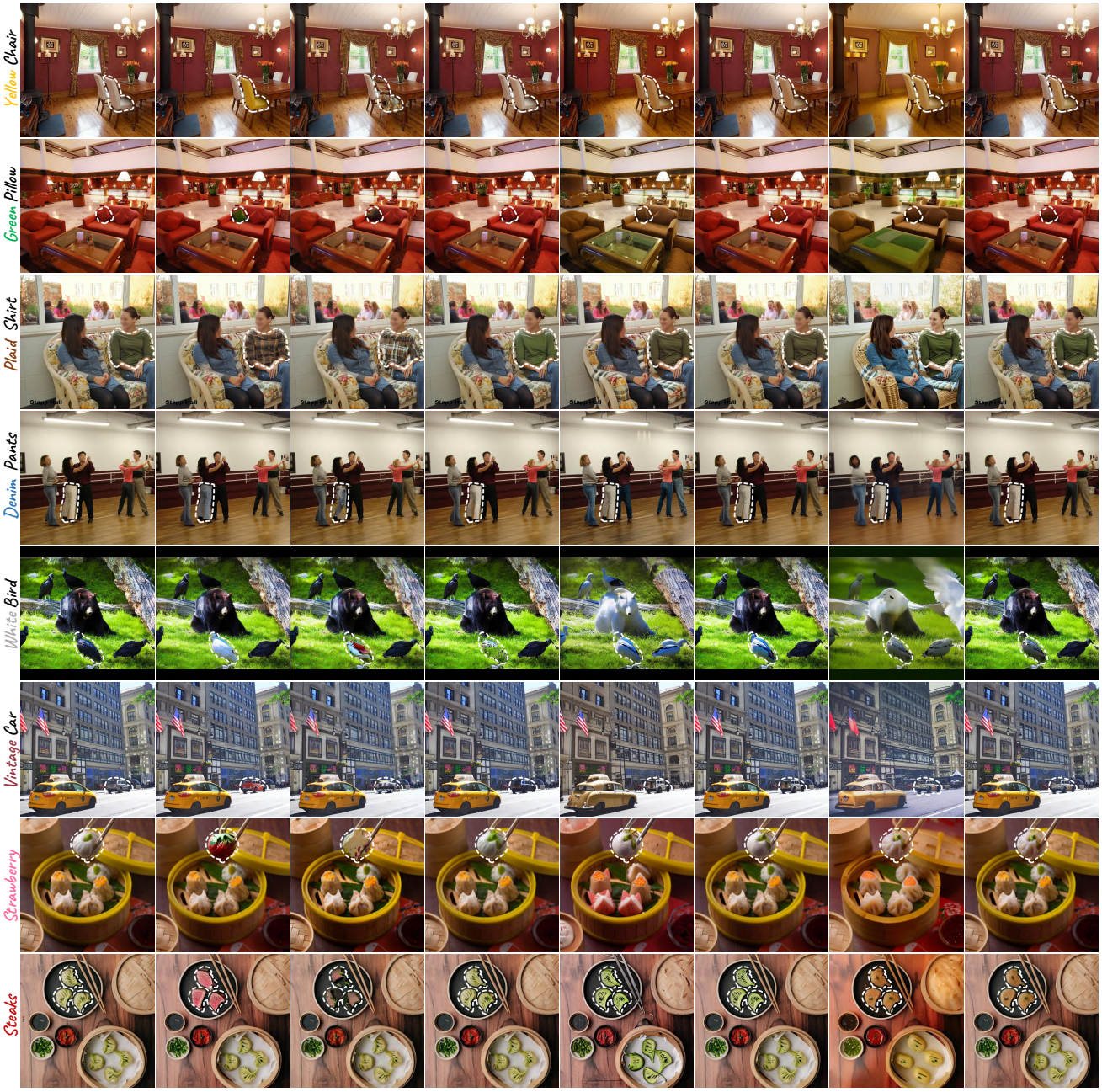
Considering that CA maps define the similarity between the input features and text embeddings, larger CA values indicate better alignment. This observation inspires the formulation of two mask-based constraints, aiming to maximize the CA value ratio in both token and spatial aspects within the predefined editing region. To illustrate, first consider the CA map $(A_t^*)_i$ of a new editing token s_i^* within a specific mask region $\mathcal{M}$ such as "**blue**" in Fig. 3 (a).

Token Ratio Constraint. Since CA maps of common tokens from the reconstruction process are first injected into the editing branch, this leads to the CA value of the new token $(A_t^*)_i$ being comparatively lower in contrast to other common tokens. We then introduce a token ratio constraint that prioritizes increasing the value proportion of the new token within mask $\mathcal{M}$ among all tokens:

$$\mathcal{L}_{TR} = \left(1 - \frac{1}{\overline{\mathcal{M}}} \sum \mathcal{M} \odot \frac{(A_t^*)_i}{(A_t^*)_i + \sum_{j=1}^{J} (A_t)_j}\right)^2, \quad (5)$$

where $\overline{\mathcal{M}}$ represents the total number of elements within the mask.

Spatial Ratio Constraint. In scenarios demanding significant editing granularity, the token ratio constraint might not sufficiently amplify the CA value $(A_t^*)_i$ within $\mathcal{M}$. To address this limitation, we introduce an additional spatial formulation, which is designed to maximize the CA values



(a) Source Image **(b) Ours** (c) Blended LD (d) DiffEdit (e) P2P (f) P2P+Blend (g) PnP (h) PnP +Blend

Figure 4. **Qualitative comparisons of localized image editing across various complex scenarios.** We highlight the editing regions with white dashed lines. Simplified target edit prompts are denoted on the left side of (a) source images. Our proposed method (b) not only achieves superior editing effects but also better preserves the structure in local regions against other baselines (c-h).

within the masked region while simultaneously minimizing them outside the mask as,

$$\mathcal{L}_{SR} = \lambda \underbrace{\left(1 - \frac{\sum \mathcal{M} \odot (A_t^*)_i}{\sum (A_t^*)_i}\right)}_{\text{Out-mask}} - \underbrace{\frac{\sum \mathcal{M} \odot (A_t^*)_i}{\sum (A_t^*)_i}}_{\text{In-mask}}, \tag{6}$$

where λ is a balance weight, and we set $\lambda = 3$ empirically.

Negative Prompt Constraint. In real image editing, the latent noise feature z_T derived by the inversion methods still retains information related to the original image $\mathcal{I}$. Achieving the desired editing results can be challenging in some cases when there is a significant difference between the texture in the original image and modified prompt $\mathcal{P}^*$, such as transferring color from "black" to "white". Our proposed

Algorithm 1: A Denoising Step using MAG-Edit

Input: A original and edited prompt $\mathcal{P}$, $\mathcal{P}^*$; a timestep t and corresponding noise latent features of reconstruction and editing branches z_t, z_t^*; a maximum iteration step MAX_IT; a function $\mathbf{F}(\cdot)$ for computing proposed constraints;a pre-trained Stable Diffusion model SD.

Output: the noisy latent feature z_{t-1}^* for the next timestep of the editing branch.

for $i = 1$ **to** MAX_IT **do**
  $\_, A_t \leftarrow SD(z_t, \mathcal{P}, t)$;
  $\_, A_t^* \leftarrow SD(z_t^*, \mathcal{P}^*, t)$;
  $\hat{A}_t \leftarrow Inject(A_t, A_t^*)$;
  $\mathcal{L} \leftarrow \mathbf{F}(\hat{A}_t)$;
  $z_t^* = \mathcal{M} \odot (z_t^* - \delta \nabla_{z_t^*} \mathcal{L}) + (1 - \mathcal{M}) \odot z_t^*$;
end
$\_, A_t \leftarrow SD(z_t, \mathcal{P}, t)$;
$\_, A_t^* \leftarrow SD(z_t^*, \mathcal{P}^*, t)$;
$\hat{A}_t \leftarrow Inject(A_t, A_t^*)$;
$z_{t-1}^* \leftarrow SD(z_t^*, \mathcal{P}^*, t)\{\hat{A}_t\}$;
Return z_{t-1}^*

method can also be used to attenuate the textural information associated with the original image $\mathcal{I}$ by employing negative prompts. In particular, we define a set of negative tokens $\mathcal{S}_{\text{ng}}^*$ to present the texture of $\mathcal{I}$ in contrast to the new tokens $\mathcal{S}^*$. For example, if $\mathcal{P}^*$ is "a man wears a white T-shirt" and the T-shirt in $\mathcal{I}$ is black, then the negative token would be "black". Consequently, we can establish the negative prompt constraint $\mathcal{L}_{\text{ng}}$ using the negative token's corresponding CA value and optimize the noise latent feature in the opposite direction as follows,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{p}} \mathcal{L} - \lambda_{\text{ng}} \mathcal{L}_{\text{ng}}, \tag{7}$$

where λ_{p} and λ_{ng} aim to balance between positive and negative prompt constraint.

4.3. Perform Gradient Guidance

Upon establishing the mask-based constraints, we compute their gradients to determine the optimal direction for modifying the current noise latent feature z_t^*. In particular, to restrict the editing effect to the predefined region, we update the noise latent feature z_t^* inside the mask $\mathcal{M}$ using the following equation:

$$z_t^* = \mathcal{M} \odot (z_t^* - \delta \nabla_{z_t^*} \mathcal{L}) + (1 - \mathcal{M}) \odot z_t^*, \tag{8}$$

where the term δ represents the gradient update scale. As detailed in Algorithm 1, z_t^* is iteratively refined until reaching the maximum number of iteration.

Method	Quantitative Metrics		Human Preference (Ours vs.)		
	CLIP Score (↑)	DINO-ViT Distance (↓)	Text Alignment (%)	Structure Preservation (%)	Overall Preference (%)
Blended LD [2]	19.12	0.089	84 %	**75%**	80%
Diffedit [8]	19.20	0.083	77 %	66%	71 %
P2P [10]	20.02	**0.079**	**87 %**	54 %	**81%**
PnP [28]	19.90	0.083	**87 %**	59 %	79 %
P2P+Blend	19.77	0.081	83 %	62 %	73%
PnP+Blend	19.47	0.080	82 %	56 %	69%
Ours	**21.79**	0.081	/	/	/

Table 1. **Quantitative comparisons of localized image editing.** We assess all the metrics and human preferences in the **localized editing regions**. "Ours vs. " indicates the proportion of users who favor our proposed method over the comparative approach.

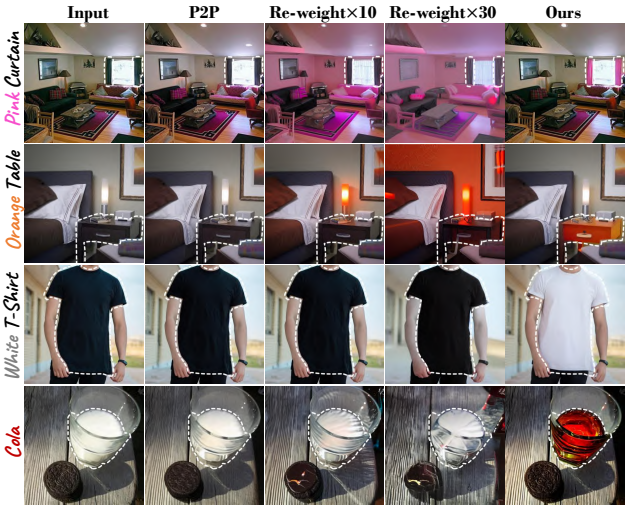


Figure 5. **Attention re-weighting [10] vs. Ours.** *Attention re-weighting* [10] either amplifies the entire editing magnitude in incorrect regions (first two rows) or fails to edit regions that significantly contradict the target prompt (last two rows). In contrast, our proposed method effectively addresses both scenarios.

Moreover, our proposed method can be readily adapted for multiple prompt editing as:

$$z_t^* = \mathcal{M} \odot (z_t^* - \delta \nabla_{z_t^*} \sum_{i=1}^{I} (\lambda_1 \mathcal{L}^1 + \cdots + \lambda_i \mathcal{L}^i + \lambda_I \mathcal{L}^I)) + (1 - \mathcal{M}) \odot z_t^*, \tag{9}$$

where the term λ_* controls the editing granularity of each prompt, with their sum equaling 1. Fig. 8 (a) demonstrates how our proposed method effectively balances the editing granularity for various prompts.

5. Experiments

5.1. Implementation details

We adopt the pre-trained Stable Diffusion v1.4 [23] model as the backbone. All CA values are calculated in the resolution of 16×16 of the U-Net, which is known to process the most semantically rich information [5]. To preserve the original information in the regions outside the mask, we incorporate latent blend operation in [10]. In practice, we

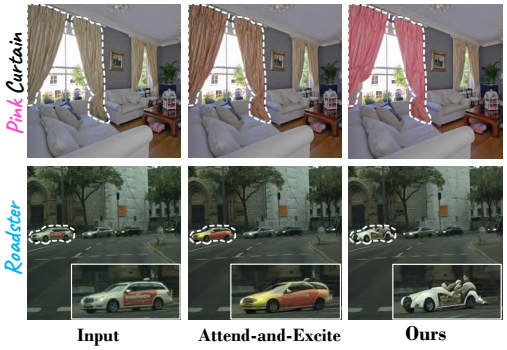


Figure 6. **Attend-and-Excite [5] vs. Ours.** Attend-and-Excite's CA constraint [5] yields unsatisfactory editing results.

select between $\mathcal{L}_{TR}$ and $\mathcal{L}_{SR}$, depending on the required granularity of the edit, allowing for adaptability across various editing types. All experiments are conducted on a single NVIDIA A100 GPU. Additional implementation details are provided in Appendix B.

5.2. Comparisons with Baselines

Benchmark Dataset. Existing datasets for text-based image editing methods primarily focused on relatively simple scenes dominated by prominent objects. To enable a more comprehensive evaluation of our method, we have curated a benchmark data set consisting of 200 images sourced from ADE20K [33], MS-COCO [16], Cityscape [7], and the Internet. The selected images feature complex compositions with multiple objects in a wide range of real-world indoor and outdoor scenes. Our evaluation primarily targets localized editing of color, texture, and object replacement. We generate the source and target prompts using GPT-4 [19]. The corresponding edit masks are obtained using the Segment Anything method[1]. Consequently, each image in the dataset is associated with three annotations: a source image prompt, a target image prompt, and the editing mask. Additional details are provided in Appendix C.1.

Baselines. We conduct comparisons with existing representative training-free diffusion-based image editing methods, covering these categories:

- **Mask-based**: Blended Latent Diffusion (Blended LD) [2] and DiffEdit [8].
- **Mask-free**: P2P [10] and PnP [28].
- **Mask-free with blending**: We combine blending operations with P2P and PnP as baselines, denoted as P2P+Blend and PnP+Blend, respectively. Note that P2P+Blend can be viewed as the proposed method w/o MAG.

Qualitative results. Fig. 4 clearly shows that Blended LD [2] leads to considerable structural changes in complex

[1] https://github.com/facebookresearch/segment-anything

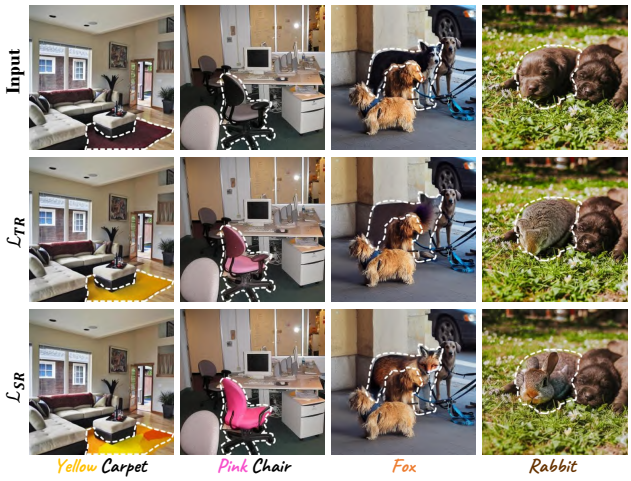


Figure 7. **Editing granularity of proposed constraints.** The token ratio constraint $\mathcal{L}_{TR}$ efficiently preserves the inherent structure in the edited region, while the spatial ratio constraint $\mathcal{L}_{SR}$ enhances editing granularity.

scenarios, resulting in significant discordance with the surrounding context. Meanwhile, DiffEdit [8], which employs DDIM inversion for foreground generation, either alters the structure, as seen in the "white bird" example in the fifth row, or fails to produce a noticeable editing effect in the intended region, as is apparent in other images. Concerning mask-free methods, P2P [10] and PnP [28] often exhibit leakage into adjacent regions, leading to minimal effects in the prospective region. This issue is particularly noticeable in tasks such as changing the color of a yellow chair or a green pillow. Blending operations might reduce leakage in some scenarios, yet the issue of misalignment persists, resulting in inefficiencies in the intended edit regions. In contrast, our proposed method shows improved editing performance with better structural preservation.

Quantitative results. We quantitatively evaluate our proposed method against baseline models using both automatic metrics and human evaluations.

Automatic Metrics. To better evaluate localized editing ability, we use the bounding boxes to crop the editing regions [12] and evaluate the image-text alignment and structure preservation using the CLIP score [21] and the DINO-ViT self-similarity distance [27], respectively. Table 1 illustrates that our proposed method significantly enhances text alignment within local regions, achieving much higher local CLIP values without compromising fidelity.

User study. We perform a user preference evaluation via pairwise comparisons on Amazon MTurk[2], focusing on text alignment, structure preservation, and overall preference in localized editing regions. As shown in Table 1, the percentages represent the proportion of users who prefer our

[2] https://www.mturk.com/

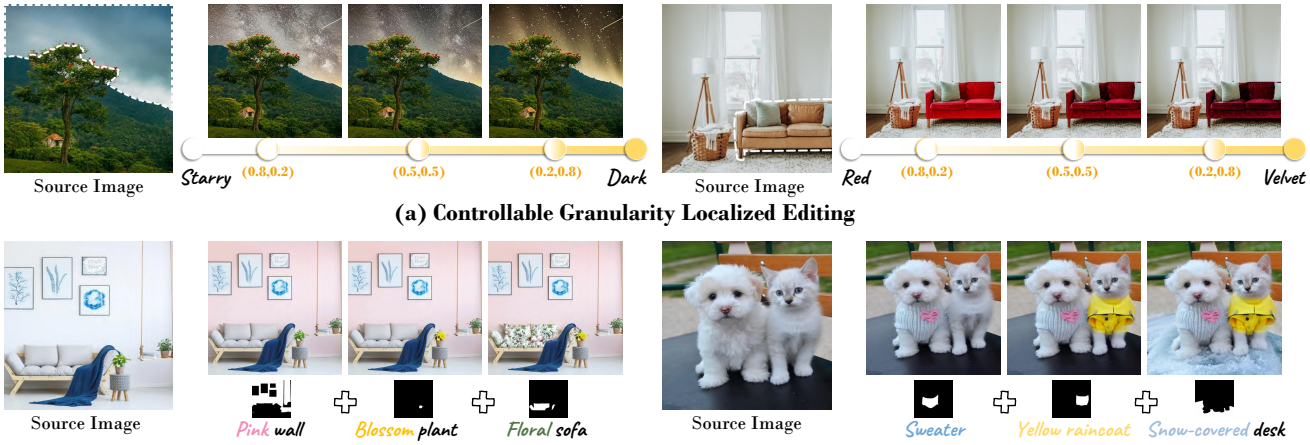


Figure 8. **Other localized editing applications of the proposed MAG-Edit.**

proposed method over comparative approaches. A significant majority, ranging from 77% to 87%, believe that our method achieves much better text alignment compared to other methods. Furthermore, our method is preferred for better structure preservation by 75% of users over Blended LD [2]. Due to its more effective balance between editability and fidelity, our proposed method is overall favored by 69% to 80% of the participants.

5.3. Ablation Study

Why Optimize the z_t. To enhance the editing effect, a straightforward approach is to directly increase the CA values of the corresponding token using the *attention reweighting* in P2P [10]. However, due to P2P's inherent misalignment, this direct amplification of CA values tends to intensify the editing effects in incorrect regions, failing to enhance the desired localized areas (first two rows in Fig. 5). Additionally, minimizing the influence of information from the original image that conflicts with the edit prompt poses a challenge, even in prominent objects, such as changing color from "black" to "white" (last two rows in Fig. 5). In contrast, our proposed method focuses on local alignment in specific regions and effectively attenuates contradicting information by directly optimizing the noise latent feature.

vs. Attend-and-Excite. Attend-and-Excite [5] optimizes the noise latent feature z_t in the unconditional generation to maximize the largest CA value of the subject token. We then establish a baseline using the CA constraint formulation in [5]. However, Fig. 6 shows that this constraint is insufficient for image editing scenarios. Contrary to unconditional generation, the noise latent feature derived from inversion methods contains more information related to the real image, as opposed to random noise features sampled from a Gaussian distribution. As such, our proposed constraints offer more efficient guidance for adjusting the noise latent feature, thereby more aptly addressing the needs of image editing.

Impact of Proposed Constraints. $\mathcal{L}_{TR}$ and $\mathcal{L}_{SR}$ offer distinct levels of editing granularity, as demonstrated in Fig. 7. $\mathcal{L}_{TR}$ excels in maintaining the inherent structure within the edit region, which aids in achieving natural color and texture modifications. On the other hand, $\mathcal{L}_{SR}$ provides stronger guidance by directly amplifying the CA values within the mask, leading to more noticeable structural changes in the edit region. As a result, $\mathcal{L}_{SR}$ is better suited for edits involving large structural shape changes.

5.4. Other Applications

MAG-Edit is also adaptable for controllable granularity and iterative localized editing. In Fig. 8 (a), MAG-Edit demonstrates the ability to balance editing granularity across various prompts, catering to user-specific requirements. Furthermore, Fig. 8 (b) illustrates MAG-Edit's capability to execute iterative, localized manipulations on various objects within a single image. More results are shown in Appendix F.

6. Conclusions

In this work, we introduce a novel technique *Mask-Based Attention-Adjusted Guidance* (MAG-Edit), specifically crafted for localized editing in complex scenarios. In particular, we propose to maximize two mask-based CA constraints, namely token and spatial ratio, to locally optimize the noise latent feature for enhanced alignment with the target text embedding. Our experimental results, both quantitative and qualitative, consistently illustrate that MAG-Edit outperforms existing methods in localized image editing within complex scenarios. We believe the proposed MAG-Edit scheme has pioneered a novel direction for applying localized editing in real-world scenarios.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 2, 3

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM TOG*, pages 1–11, 2023. 1, 2, 3, 6, 7, 8, 11

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 2, 3, 13, 14

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023. 2, 3, 17

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*, pages 1–24, 2023. 3, 6, 7, 8

[6] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 3, 11

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 7, 11

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, pages 1–22, 2023. 1, 2, 3, 6, 7, 11

[9] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Yuxiao Chen, Di Liu, Qilong Zhangli, et al. Improving negative-prompt inversion via proximal guidance. *arXiv preprint arXiv:2306.05414*, 2023. 16

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, pages 1–36, 2023. 1, 2, 3, 4, 6, 7, 8, 11, 13, 16

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeruIPS workshop*, pages 1–14, 2021. 3

[12] Wenjing Huang, Shikui Tu, and Lei Xu. Pfb-diff: Progressive feature blending diffusion for text-driven image editing. *arXiv preprint arXiv:2306.16894*, 2023. 2, 3, 7, 12

[13] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 16

[14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 2, 3

[15] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023. 3, 16

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 7, 11

[17] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 3

[18] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 2, 3, 4, 11

[19] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7, 11

[20] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, pages 1–11, 2023. 2, 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 7

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 6

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2, 3

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 2

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, pages 1–20, 2021. 3, 11

[27] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *CVPR*, pages 10748–10757, 2022. 7

[28] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 1, 2, 3, 6, 7

[29] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*, 2023. 2, 3

[30] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, pages 7452–7461, 2023. 3

[31] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023. 2, 3, 13, 14

[32] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, pages 6027–6037, 2023. 2, 3, 14

[33] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 7, 11

Appendix

A. Summary

In this appendix, we present more implementation details, additional experiments, and additional results as follows:

- We present more implementation details of MAG-Edit in Section B. Furthermore, Section C illustrates more implementation details on the benchmark dataset, baselines, quantitative metrics, and user study.
- In Section D, we extend our comparisons to encompass training and fine-tuning methods, as well as recent developments in inversion techniques. For a more comprehensive understanding of our proposed method, additional ablation studies are conducted and detailed in Section E.
- We demonstrate additional qualitative results to complement the paper in Section F.
- Finally, the limitations of our approach are thoroughly analyzed in Section G.

B. Implementation Details

We utilize the official pre-trained Stable Diffusion v1.4 model[3] as our foundation model. The denoising sampling process employs the DDIM method [26] over $T = 50$ steps, maintaining a constant classifier-free guidance scale of 7.5. CA injection is performed during $[T, \tau_1]$. For varying editing requirements, we set $\tau_1 = 10$ for color and texture edits, and $\tau_1 = 40$ for shape variation edits. Our **MAG-Edit** optimization takes place during diffusion steps in the range $[T, \tau_2]$, with τ_2 empirically set to 25. For the gradient guidance process, we follow [6] by setting the gradient update scale δ using a linear scheduling rate as $\sqrt{(1-\alpha_t)/\alpha_t}$, particularly to optimize the token ratio constraint $\mathcal{L}_{TR}$. This approach modulates the gradient's magnitude based on the denoising progress. On the contrary, for the constraint of the spatial ratio $\mathcal{L}_{SR}$, we keep $\delta = 1$. The optimization process is also influenced by the maximum number of iterations, empirically set $\mathrm{MAX\_IT} = 15$. In cases involving negative prompt constraints, we empirically set $\lambda_p = 2.5$ and $\lambda_{ng} = 5.5$. To further preserve the structure of the original image, we also consider incorporating self-attention as P2P [10] and replace them at diffusion steps $t \in [T, 25]$. Towards the end of the denoising process $t \in [15, 0]$, we implement a latent blend operation from P2P [10] to maintain information outside the edited region mask $\mathcal{M}$. When evaluated on an Nvidia A100 (40GB) GPU, the runtime of MAG-Edit is around $1 \sim 5$ minutes, varying with the selected values of $\mathrm{MAX\_IT}$ and τ_2.

[3] https://github.com/CompVis/stable-diffusion

C. Details of Comparisons with Baselines

C.1. Benchmark Dataset

Current datasets for text-based image editing methods are primarily limited to simple scenes with prominent objects. To enable a more thorough evaluation of our method, we have developed a benchmark dataset, named MAG-Bench, consisting of 200 images sourced from ADE20K [33], MS-COCO [16], Cityscape [7], and the Internet. This dataset features complex scenes with multiple objects in various real-world indoor and outdoor settings, encompassing a wide range of object categories like humans, furniture, animals, vehicles, and food. MAG-Bench is specifically designed to assess three types of local editing: (1) color editing, (2) texture editing which includes changes in material, background, and style, and (3) object replacement. For the generation of source and target prompts, we initially utilized GPT-4 [19], followed by manual refinement to ensure the accuracy and relevance of these prompts. The corresponding editing masks for each image are derived using the Segment Anything method[4]. Acknowledging the critical role of the mask's size in localized editing, we initially classify each image into three categories based on mask size: relatively small, medium, and relatively large. We then ensure a balanced distribution of varying sizes of editing regions across the datasets. Thus, each image in MAG-Bench is accompanied by three annotations: a source prompt, a target edit prompt, and an edit region mask, as illustrated in Fig. 9.

C.2. Implementation Details of Baselines

We use the official codes released by the authors for Blended LD[5], P2P[6], and PnP[7]. For DiffEdit [8], we adopt the implementation from InstructEdit[8], which enhances automatic mask generation for scenarios involving multiple objects. This implementation, while improving upon mask generation, does not modify the core editing algorithm of DiffEdit [8]. To facilitate fair comparisons, all methods use *identical masks* provided in our benchmark dataset. Notably, for DiffEdit [8] and P2P [10], we utilize ground-truth masks instead of those generated through unsupervised learning or derived from average CA maps. In the case of P2P [10], we also integrate Null-text inversion [18] as our approach for encoding real images. With the exception of Blended LD [2], which solely focuses on the target edit description for the foreground region and omits tokens for other unedited areas, all other methods employ target prompts identical to those used in our method.

[4] https://github.com/facebookresearch/segment-anything

[5] https://github.com/omriav/blended-latent-diffusion

[6] https://github.com/google/prompt-to-prompt

[7] https://github.com/MichalGeyer/plug-and-play

[8] https://github.com/QianWangX/InstructEdit

Scenarios	Edit Type	Source Image	Source Prompt	Target Prompt	Mask	Mask Type
Outdoor	Color		A couple and a kid with **black** hair are sitting on the bench	a couple and a kid with **blond** hair are sitting on the bench		Small
	Object		A green **truck** and some cars park under a tall building.	A green **bus** and some cars park under a tall building.		Medium
	Texture		Guinea fowl stand on **dry grass** under sky.	Guinea fowl stand on **desert** under sky.		Large
Indoor	Color		The wooden and round table is surrounded by four wooden chairs and a light **brown** chair is next to the windows.	The wooden and round table is surrounded by four wooden chairs and a light **red** chair is next to the windows.		Small
	Object		There are a **box** and lemons and several lemons on white sheet.	There are a **bowl** and lemons and several lemons on white sheet.		Medium
	Texture		There is a table with cups and four chairs on the **plaid** carpet.	There is a table with cups and four chairs on the **bohemian** carpet.		Large

Figure 9. **Examples images and annotations in the MAG-Bench dataset.**

C.3. Evaluation Details

We utilize the CLIP score with the CLIP ViT-L/14 model, as implemented in[9], and the DINO-ViT self-similarity distance, available at[10], as our evaluation metrics. To precisely evaluate localized editing, we crop the editing regions in both the source and edited images using bounding boxes as [12]. This approach enables us to specifically assess text prompt alignment within these localized regions by calculating the CLIP score on the target edited tokens with the respective cropped edited image. For instance, in a scenario where the editing objective is to alter a car's color to red, the CLIP score is computed using the phrase "red car." This calculation excludes common tokens shared between the source and target prompts and focuses solely on the cropped image depicting the edited car and the target phrase. To evaluate structure preservation within the localized editing regions, we utilize the DINO-ViT self-similarity by calculating the distance between the cropped source image and the corresponding cropped edited image.

[9] https://github.com/showlab/loveu-tgve-2023
[10] https://github.com/omerbt/Splice

C.4. Details of User Study

We conduct a user study on the Amazon MTurk platform[11]. The user study comprises over 120 tasks, each evaluated by five human evaluators, as depicted in Fig. 10. In each task, participants are presented with a source image alongside two edited images: one generated by our proposed method and the other by a randomly selected baseline method, with their presentation order shuffled. To enhance the visibility of localized editing regions, we outline the prospective edit regions with white dashed lines in each pair of comparison images and their corresponding source images, as illustrated in Fig. 10. Additionally, a simplified version of the target edit prompt was displayed beneath the comparison images. We then pose three questions for the raters to answer:

- Text Alignment: In the dashed region, which image aligns better with the "edit prompt"?
- Structure Preservation: In the dashed region, which image preserves structures more similarly to the source image?
- Overall: In the dashed region, which image performs

[11] https://requester.mturk.com

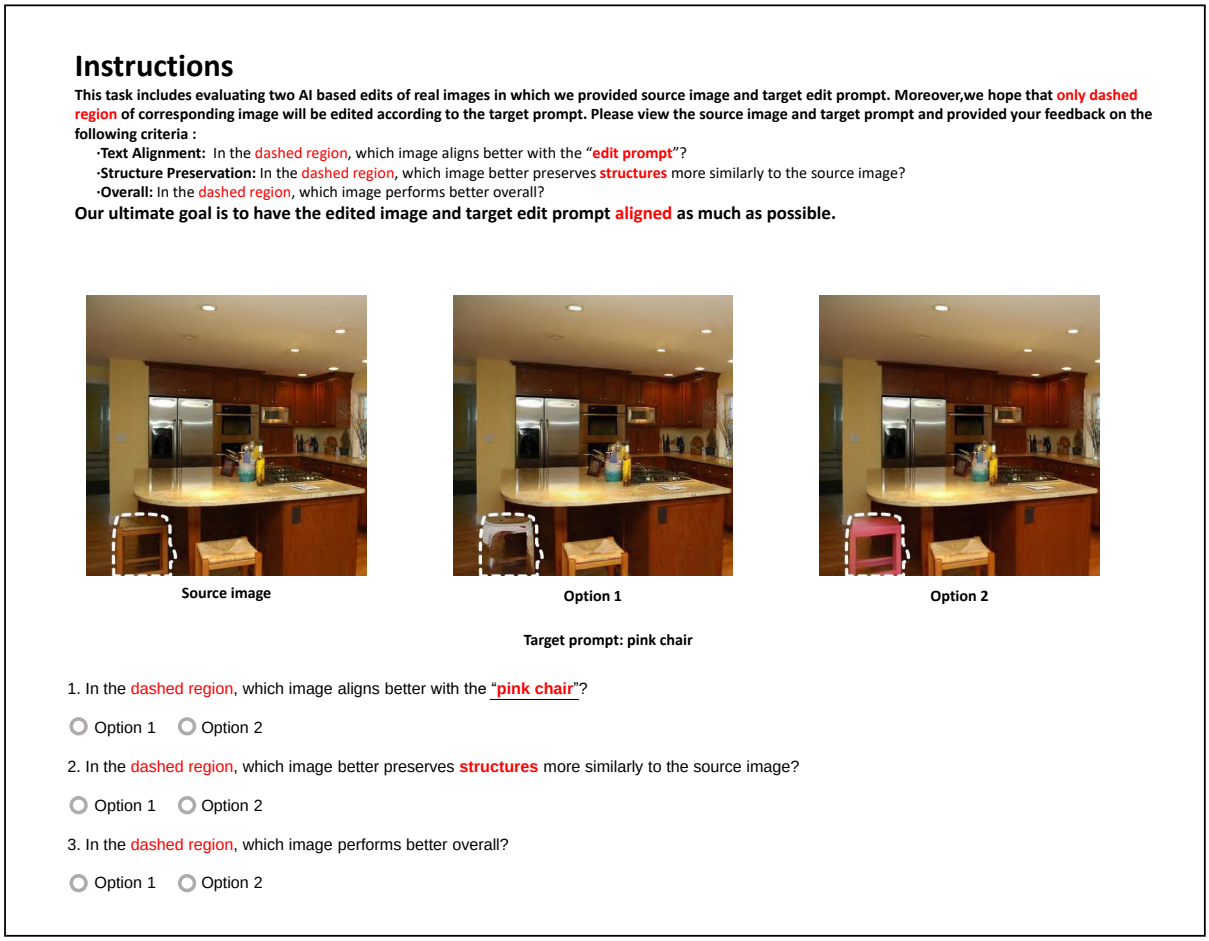


Figure 10. **Example of one task for 5 human raters on Amazon MTurk to complete.**

better overall?

To ensure the credibility and reliability of our user study, we only involve Amazon MTurk workers with 'Master' status and a Human Intelligence Task (HIT) Approval Rate exceeding 90% across all Requesters' HITs. In total, the 120 tasks garnered responses from 600 distinct human evaluators.

D. Comparisons with Other Baselines

In this section, we begin by comparing our approach with current training and fine-tuning methods, aiming to further validate the efficacy of our proposed method in facilitating localized editing in complex scenarios. Subsequently, we extend our comparison to include recent advancements in training-free inversion methods. This comparison is intended to illustrate that despite improvements in inversion methods, they still face challenges in addressing localized editing issues.

D.1. Comparisons with Training and Fine-tuning Methods

We initiate our qualitative comparison with existing training methods by evaluating InstructPix2Pix [3] and MagicBrush [31], utilizing their officially released codes and models. InstructPix2Pix [3] is trained on an extensive data set, which includes instructions generated by GPT-3 and image examples modified by P2P [10]. This training facilitates instruction-based image editing during the inference phase. MagicBrush [31] harnesses a large-scale dataset of manually annotated real image editing triplets and optimizes the InstructPix2Pix model to improve editing capabilities. For our comparisons, we utilize editing instructions such as "make" and "change" to manipulate images. Fig. 11 illustrates that InstructPix2Pix, due to its lack of mask integration, frequently leads to substantial leakage into incorrect regions during localized editing in complex scenes. In contrast, MagicBrush demonstrates better localized editing in some cases, thanks to mask-integrated examples in its dataset. However, MagicBrush encounters difficulties in

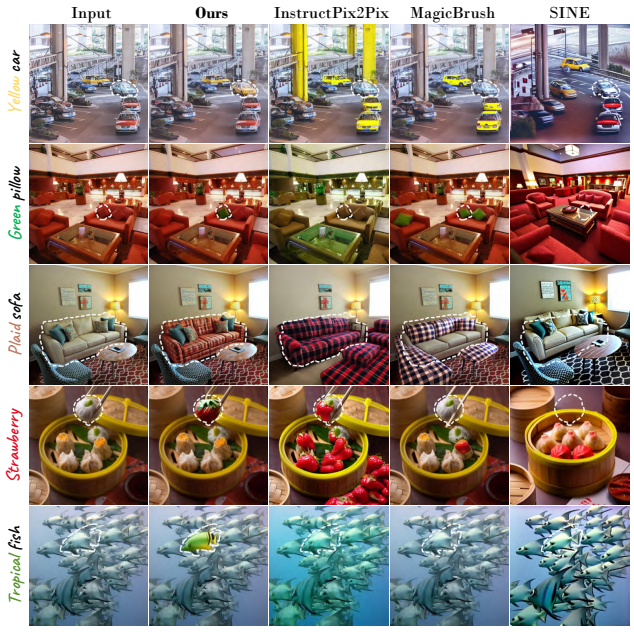


Figure 11. **Qualitative comparisons with training and fine-tuning methods for localized editing in complex scenarios.** Training approaches such as InstructPix2Pix [3] and MagicBrush [31] demonstrate issues like leakage or unintended modifications in structure. The fine-tuning method SINE [32] is ineffective in both reconstructing and generating desired editing effects.

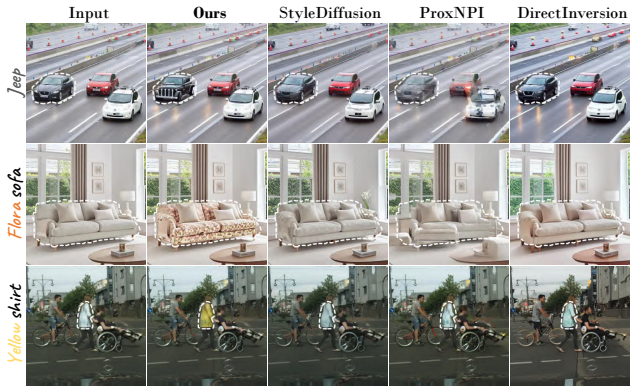


Figure 12. **Qualitative comparisons with recent inversion methods for localized editing in complex scenarios.** Despite recent advancements, solely enhancing inversion methods continues to be inadequate for effective editing of localized regions in complex scenarios.

precisely localizing individual objects within scenes containing multiple similar objects. This challenge is evident in the first and second rows of Fig. 11, where it struggles with tasks like coloring one car yellow and one pillow green. Moreover, as shown in the third row of Fig. 11, MagicBrush [31] tends to modify the underlying structure in areas undergoing texture changes. In contrast, our training-free method efficiently attains desired editing effects in the

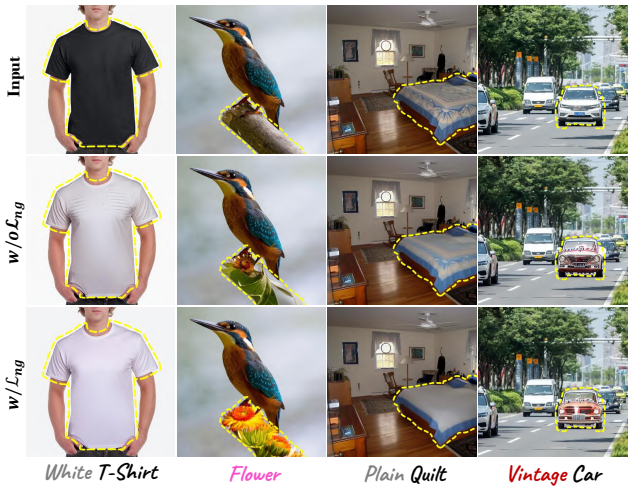


Figure 13. **Ablation study on the negative prompt constraint.** Negative prompt constraints can amplify the effectiveness of editing by diminishing the influence of information from the original image.

Figure 14. **Impact of optimization iterations.** Increasing the number of iterations enhances the granularity of editing. However, overly extensive iterations can lead to notable artifacts arising from structural modifications.

target local regions while preserving the original structure. A significant advantage of our approach is the elimination of the need for extensive training on large datasets, saving significant time and resources.

Subsequently, we compare our method with the existing fine-tuning method, SINE [32], using the code provided by its authors. SINE [32] proposes fine-tuning a pre-trained text-to-image (T2I) model with a single real image, incorporating model-based classifier guidance and patch-based guidance to prevent overfitting. However, as illustrated in Fig. 11, SINE fails to generate any noticeable editing effects

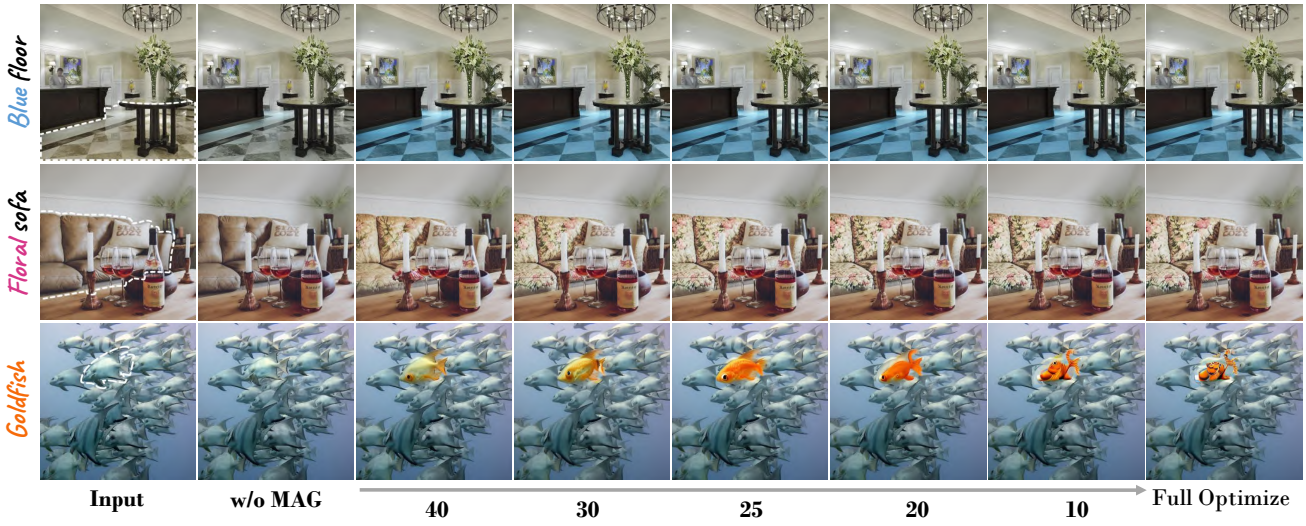


Figure 15. **Applying MAG-Edit through a varied number of diffusion steps.** We use white dashed lines to demarcate the editing regions in the source images. Each row demonstrates the optimization of the noise latent feature ranging from 0% (left) to 100% (right) of the steps. In particular, we assign values to $\tau_2 = \{50, 40, 30, 25, 20, 10, 0\}$, indicating the end of the diffusion step range, from 50 to τ_2, as noted at the bottom of each image. Without MAG, there is a negligible localized editing effect in the intended regions. On the other hand, employing MAG across all steps does not markedly enhance the granularity of color and texture editing. Moreover, it results in noticeable structural artifacts in the shape editing.

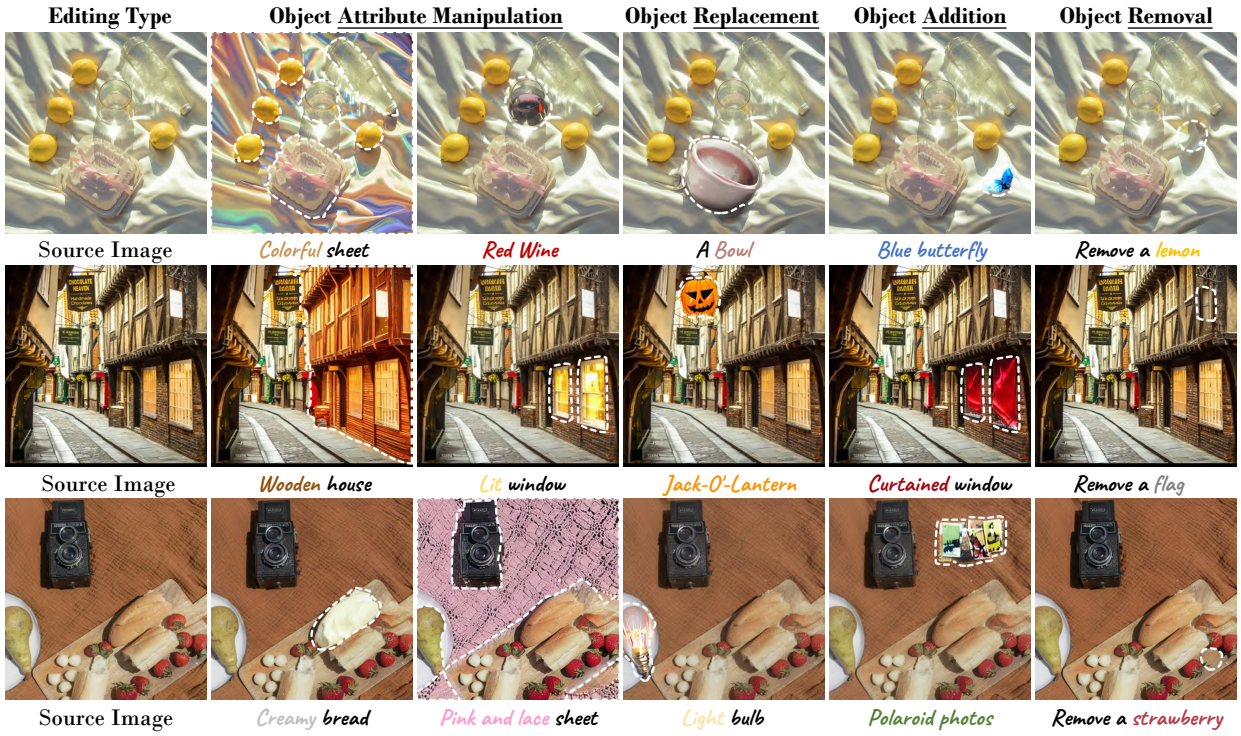


Figure 16. **Various localized editing types.** We provide a simplified version of the corresponding target prompt under each edited image.

in the intended regions. Furthermore, it faces difficulties in accurately reconstructing the original image in complex scenarios.

D.2. Comparisons with Recent Inversion Methods

To demonstrate that localized editing challenges are not sufficiently addressed by mere advancements in inversion tech-

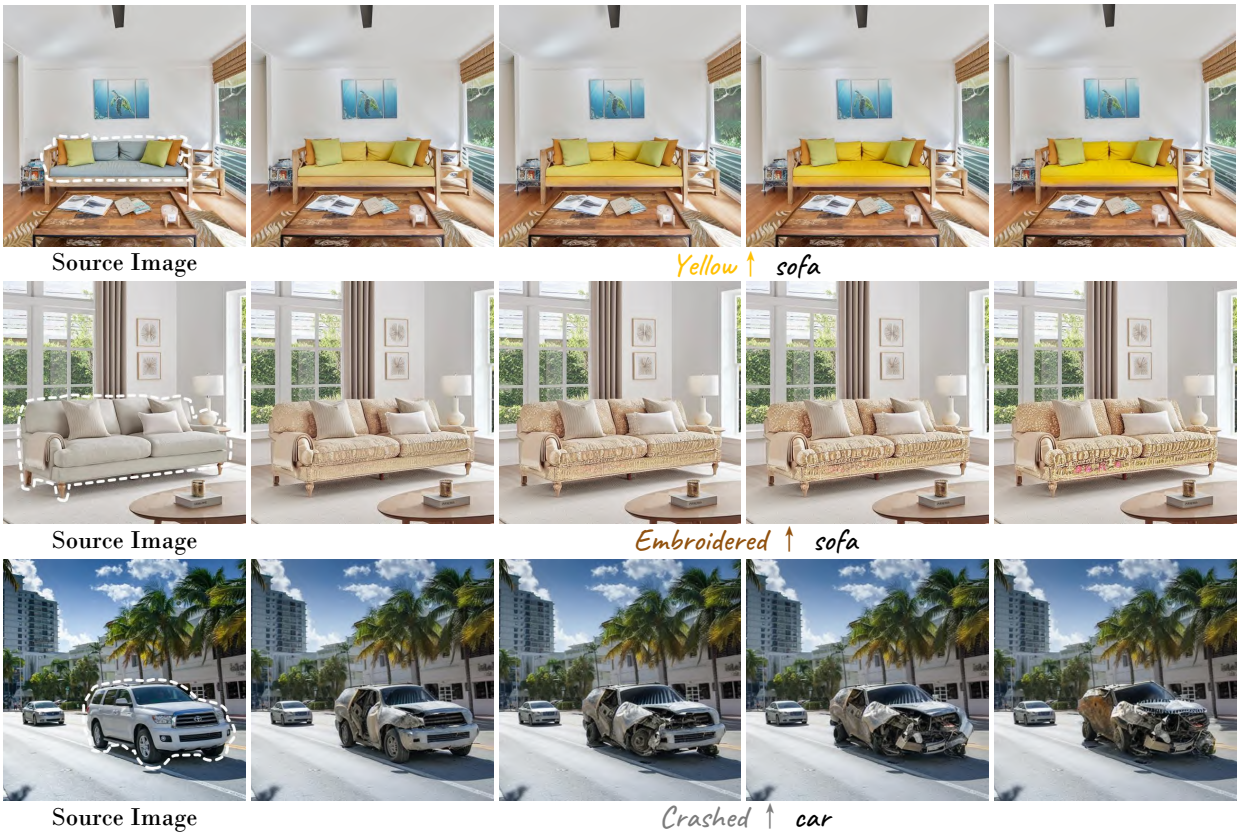


Figure 17. **Granularity controllable localized editing.** We present a simplified version of the corresponding target prompt under the edited images. ↑ denotes increasing the editing magnitude.

Figure 18. **Spatial controllable localized editing.**

niques, we compare our method with recent inversion methods. This includes Style Diffusion [15], ProxNPI [9], and DirectInversion [13], utilizing their official codes. Each of these approaches incorporates P2P [10] to facilitate editing capabilities. As depicted in Fig. 12, it is evident that these recent inversion methods are unable to produce effective editing results in localized regions within complex scenarios. This underscores the heightened challenges faced in localized editing within intricate compositions compared to simpler settings. In contrast, our method significantly improves localized editing by optimizing the noise latent feature through our specially designed MAG mechanism.

E. Additional Ablation Studies

Impact of Negative Prompt Guidance. Fig. 13 demonstrates that negative prompt guidance is effective in diminishing the original image's information, which is beneficial when dealing with original images that have information significantly contrast with the target prompt. For instance, as shown in the first column of Fig. 13, when altering the color of a T-shirt from black to white, not applying negative constraints could lead to the edited image preserving some black elements. The negative prompt constraint, in such scenarios, efficiently reduces this residual black information. Moreover, as observed in the third column of Fig. 13,

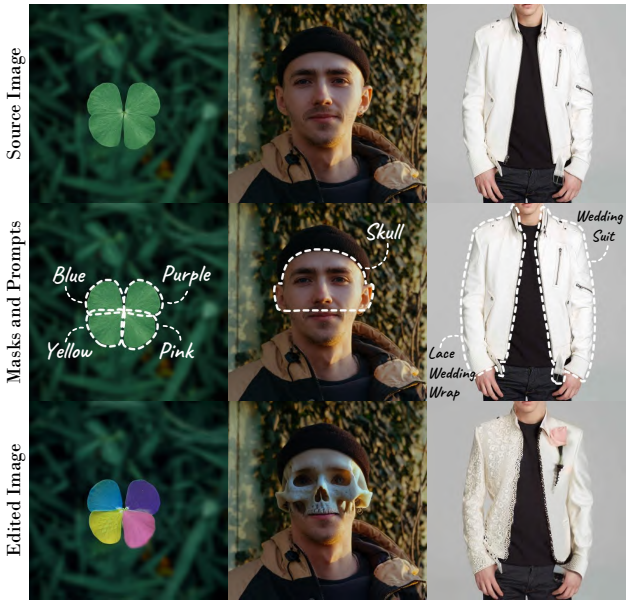


Figure 19. **Part-level localized editing.** In the second row of images, the editing regions are indicated with white dashed lines, and the target prompts are also annotated for clarity.

when transforming a patterned quilt into a plain one, the negative prompt constraint plays a crucial role in diminishing the original textures of the quilt.

Impact of Optimization Iterations. The number of maximum iterations for optimizing the noise latent feature is crucial in modulating the magnitude of editing. As shown in Fig. 14, increasing the number of iterations can improve the granularity of the editing. However, in texture and shape editing, excessive iterations may result in significant artifacts as a result of alterations in the structure.

Impact of Optimization Diffusion Steps. Applying MAG-Edit across various diffusion steps significantly impacts the final editing results. Fig. 15 demonstrates that optimization in the initial diffusion steps can quickly alter the color, indicating that optimization within the $t \in [T, 40]$ steps is generally sufficient for color editing. On the contrary, texture and shape edits necessitate a greater number of diffusion steps. Updating the latent noise feature after 25 steps does not significantly improve texture editing granularity but requires extended optimization time. In shape editing, over-optimization after 25 steps can lead to pronounced artifacts, due to structural changes.

F. Additional Results

Our method offers a broad spectrum of localized editing capabilities, encompassing object attribute manipulation (*e.g.*, color and texture), object replacement, insertion, and removal, as exemplified in Fig. 16. Additional examples

Figure 20. **Editing failure cases.** Due to its reliance on maintaining the structure using the CA maps of the reconstruction branch, the proposed method encounters limitations in editing images that necessitate significant pose alterations. For example, changing the dog from "standing" to "sitting".

of localized editing in complex scenarios are illustrated in Fig. 21 and Fig. 22. Furthermore, we demonstrate the controllability of our localized editing approach in terms of both the magnitude of edits in Fig. 17 and their spatial precision in Fig. 18. This allows for precise adjustment of editing granularity and the application of editing effects in various locations, catering to a variety of user requirements.

A key advancement of our method is *its extension from object-level to more intricate part-level localized editing*, thereby enabling the integration of various editing effects within distinct parts of a single object. As demonstrated in Fig. 19, our method is capable of sophisticated editing, such as altering a four-leaf clover into a four-colored flower, or the creation of garments with mixed textures, showcasing its versatility and precision in fine-grained localized editing tasks.

G. Limitations and Future Work

The MAG-Edit method has shown effective capabilities in localizing edits within complex scenarios, but it also has its limitations, which are a key focus of our future research efforts. A primary limitation is the method's inference time attributed to the optimization process, which takes around $1 \sim 5$ minutes on an A100 GPU to edit a single image. Future work will focus on developing strategies to accelerate this optimization process. Furthermore, our method relies heavily on maintaining structure through CA maps in the reconstruction branch. However, it falls short in editing tasks that demand substantial pose changes. As illustrated in Fig. 20, an example of this limitation is observed in the task of transitioning a standing dog to a seated position, as discussed in [4]. We acknowledge this challenge and plan to explore solutions in future work, potentially involving adjustments in how the SA is injected from the reconstruction branch to the editing branch.

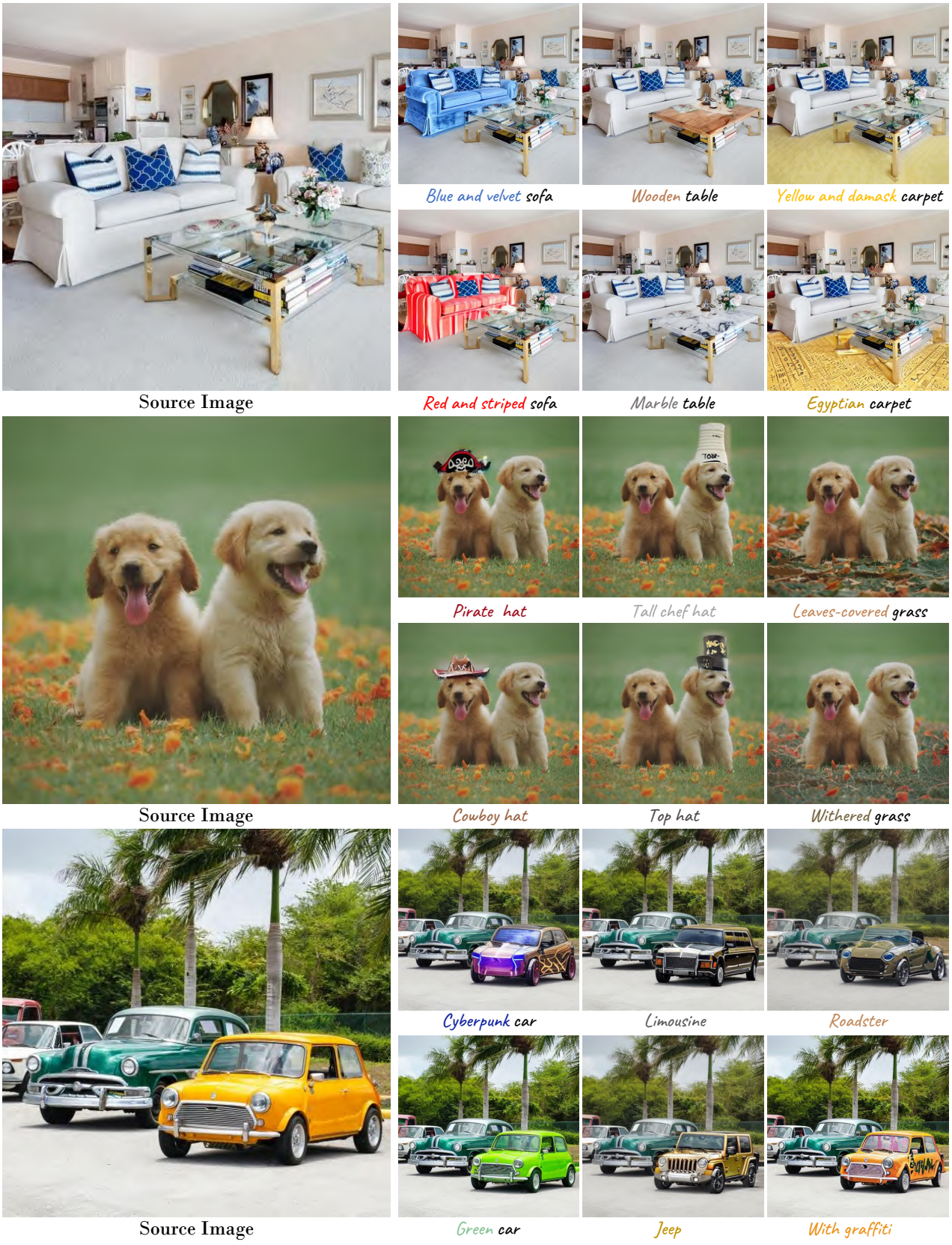


Figure 21. **Various localized editing types.** In each edited image, we present a simplified version of the corresponding target prompt.

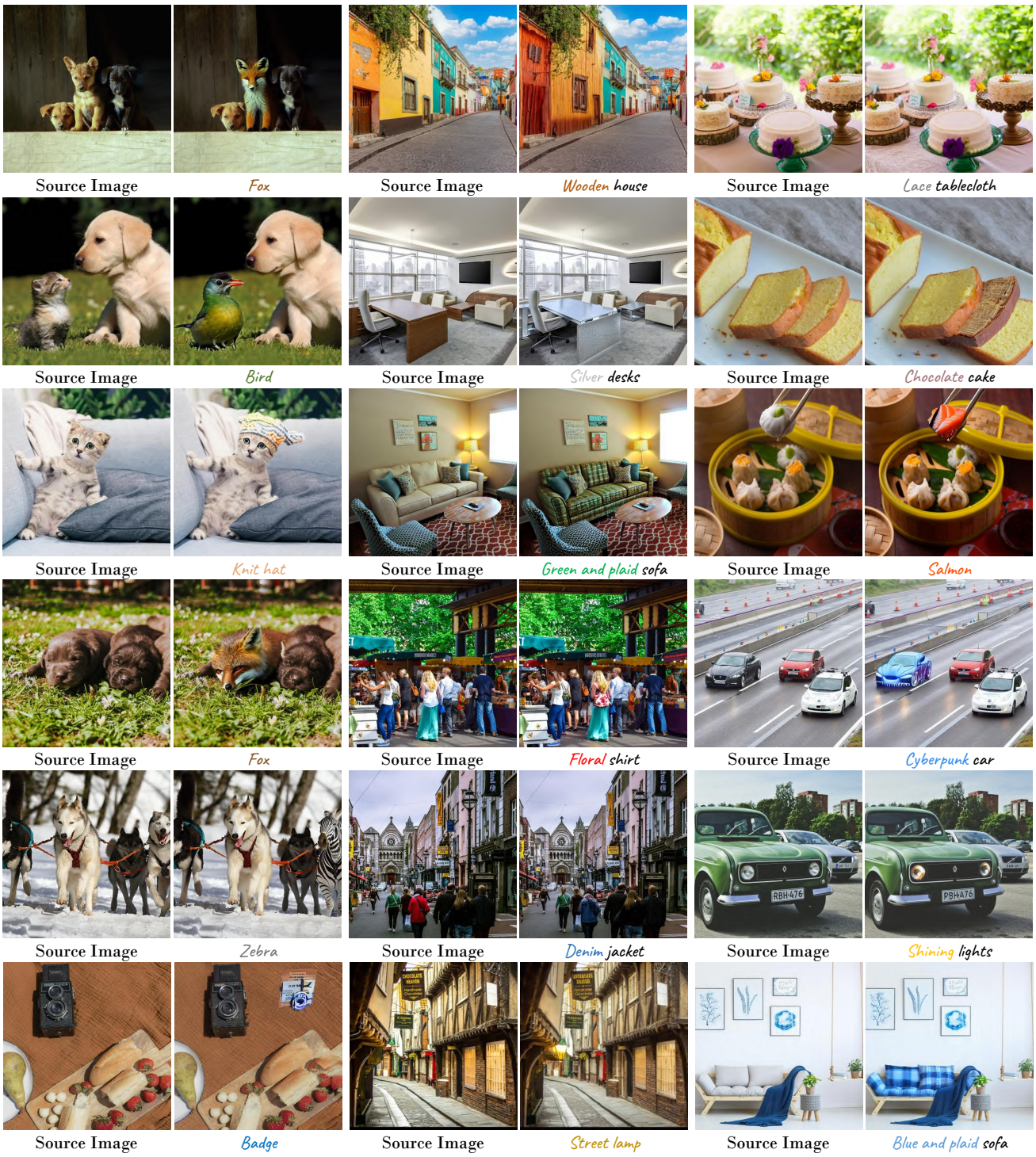


Figure 22. **Additional results on localized editing in complex scenarios.** We provide a simplified version of the target prompt beneath each edited image.